September 18, 2013

Findings of the XSEDE Cloud Use Survey

John Towns
PI and Project Director, XSEDE
Director, Collaborative eScience Programs, NCSA/Univ of Illinois
jtowns@ncsa.illinois.edu

XSEDE
Extreme Science and Engineering Discovery Environment

XSEDE Cloud Integration Investigation

- Cloud Survey Motivation
  - the goal of XSEDE is to enhance research productivity
  - XSEDE must embrace cloud
  - XSEDE must have a clear understanding of how researchers using the cloud today and why
  - based on this information XSEDE plans to integrate cloud services into its portfolio to support use cases that are not well served by its current resource offerings
- Investigation Team
  - Ian Foster, Steve Tuecke, *ANL/University of Chicago*
  - David Lifka, Susan Mehringer, Paul Redfern, *Cornell University CAC*
  - Craig Stewart, *Indiana University*
  - Manish Parashar, *Rutgers University*

XSEDE Cloud Use Survey

- XSEDE announced survey August 2012 and closed May 1, 2013
  - 80 cloud projects from around globe
    - broad participation: 21 disciplines + HASS
    - extensive technical data: 19 categories
    - user perspectives: e.g., cloud benefits/challenges
  - focused exclusively on use of cloud for research and education

- Full Web-based report will be available very soon
  - https://www.xsede.org/project-documents

Data Collected

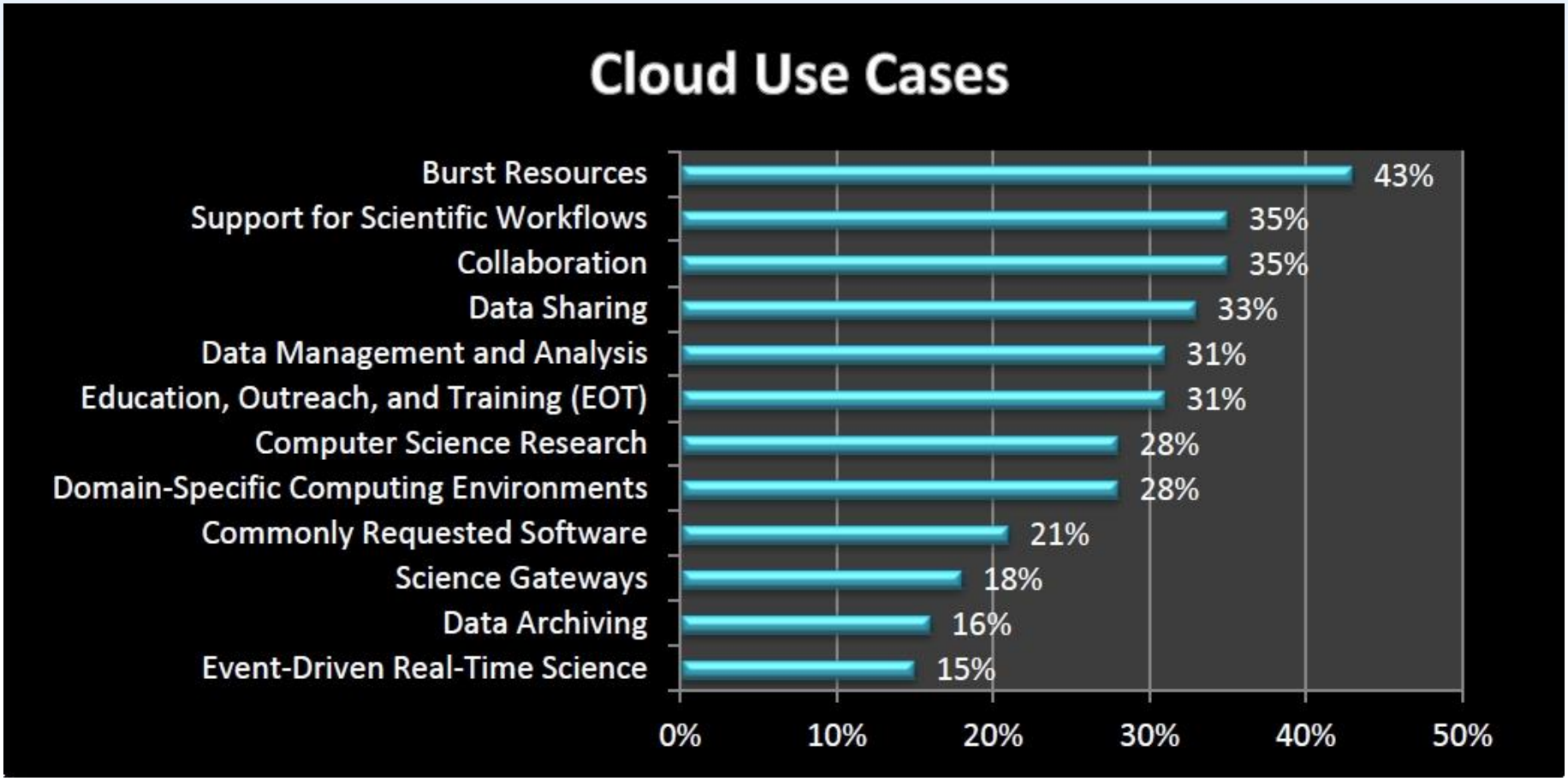
- Project Title
- Use Case
- Researchers/Investigators
- Abstract
- Cloud Providers
- Special Features
- Use Regularity
- Cores Used Peak & Cores Steady State
- Core Hours in a Year
- Access Storage
- Preferred Storage
- Amount of data Accessed During Run
- Short-Term Storage
- Long-Term Storage
- Data Moved Into Cloud
- Data Moved Out Cloud
- Bandwidth In/Out of Cloud
- Bandwidth to Storage Within
- Type Data Moving
- Data Accessed By
- Software
- Capabilities/Features
- Problems/Limitations
- Additional Notes

Representation of Disciplines

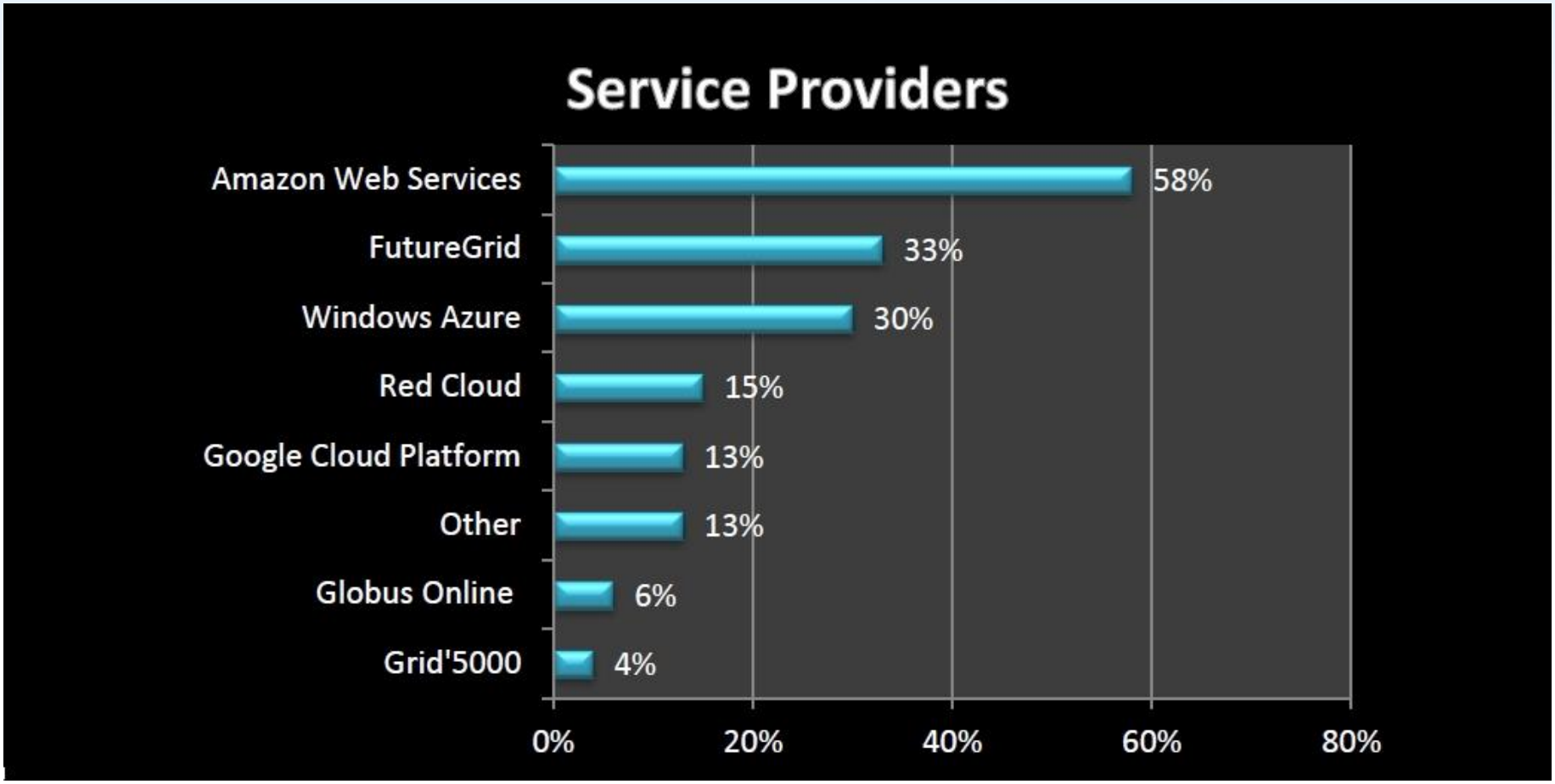
- Science & Engineering (73):
  - Astronomy (2)
  - Biology (4)
  - Biochemistry (4)
  - Biomedical Imaging Informatics (3)
  - Chemistry (2)
  - Computer Science (30)
  - Engineering (1)
  - Environmental Sciences (3)
  - Finance (1)
  - Genetics & Bioinformatics (8)
  - Geographic Information Sciences (1)
  - Geosciences (3)
  - Industrial Engineering (1)
  - Materials Science (1)
  - Neuroscience (2)
  - Operations Research (1)
  - Plant Pathology (1)
  - Physics (2)
  - Physiology & Biophysics (2)
  - Systems Engineering (1)
- Humanities, Art, and Social Sciences(5):
  - Cross-HASS Data Repository (1)
  - Economics (1)
  - Linguistics (2)
  - Social Sciences (1)
- Discipline Unspecified (2):
  - Investigations of Cloud Technologies (2)

Sample Cloud Providers

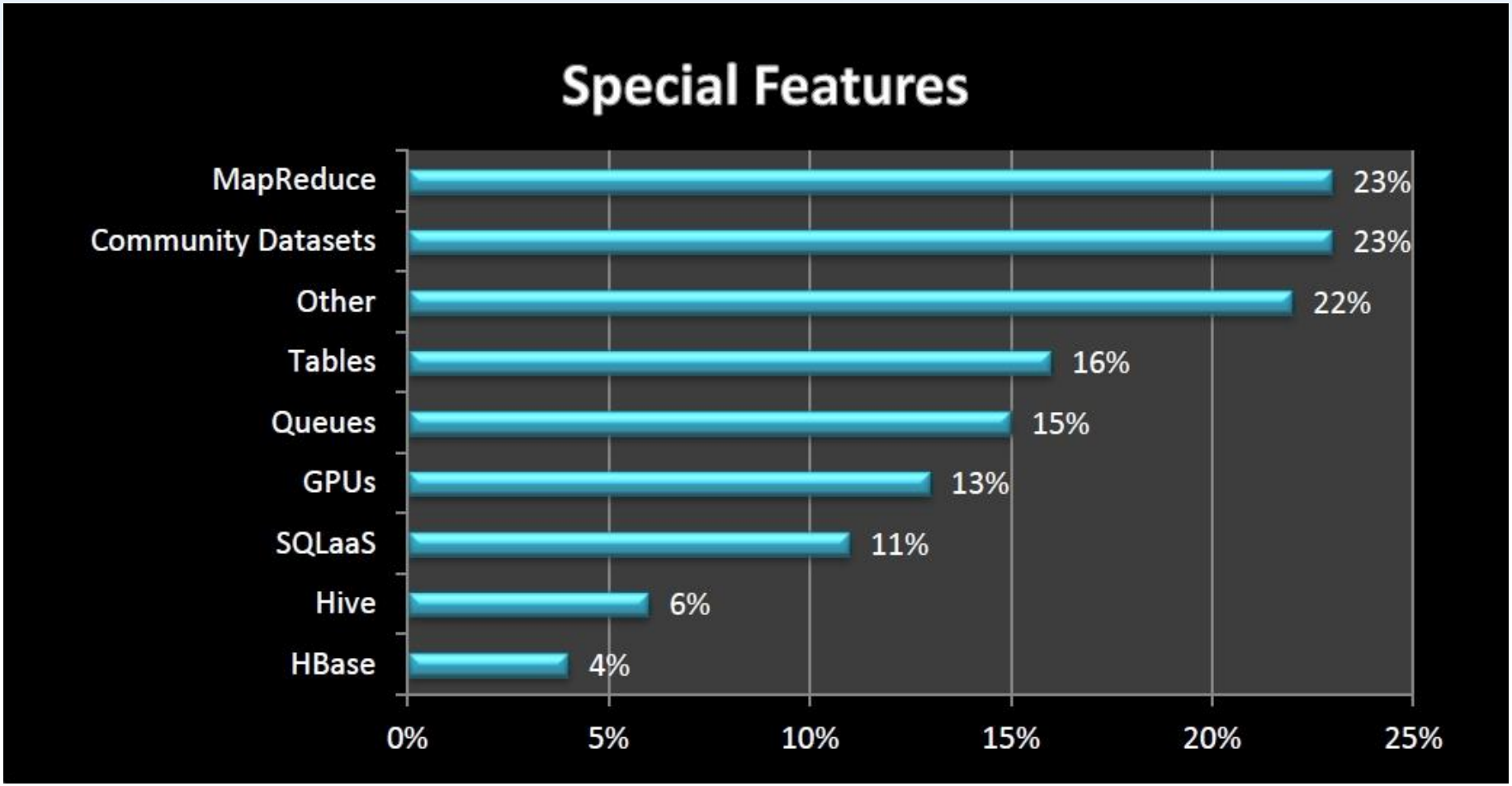
What cloud use cases are represented by your research or education project? Check all that apply.



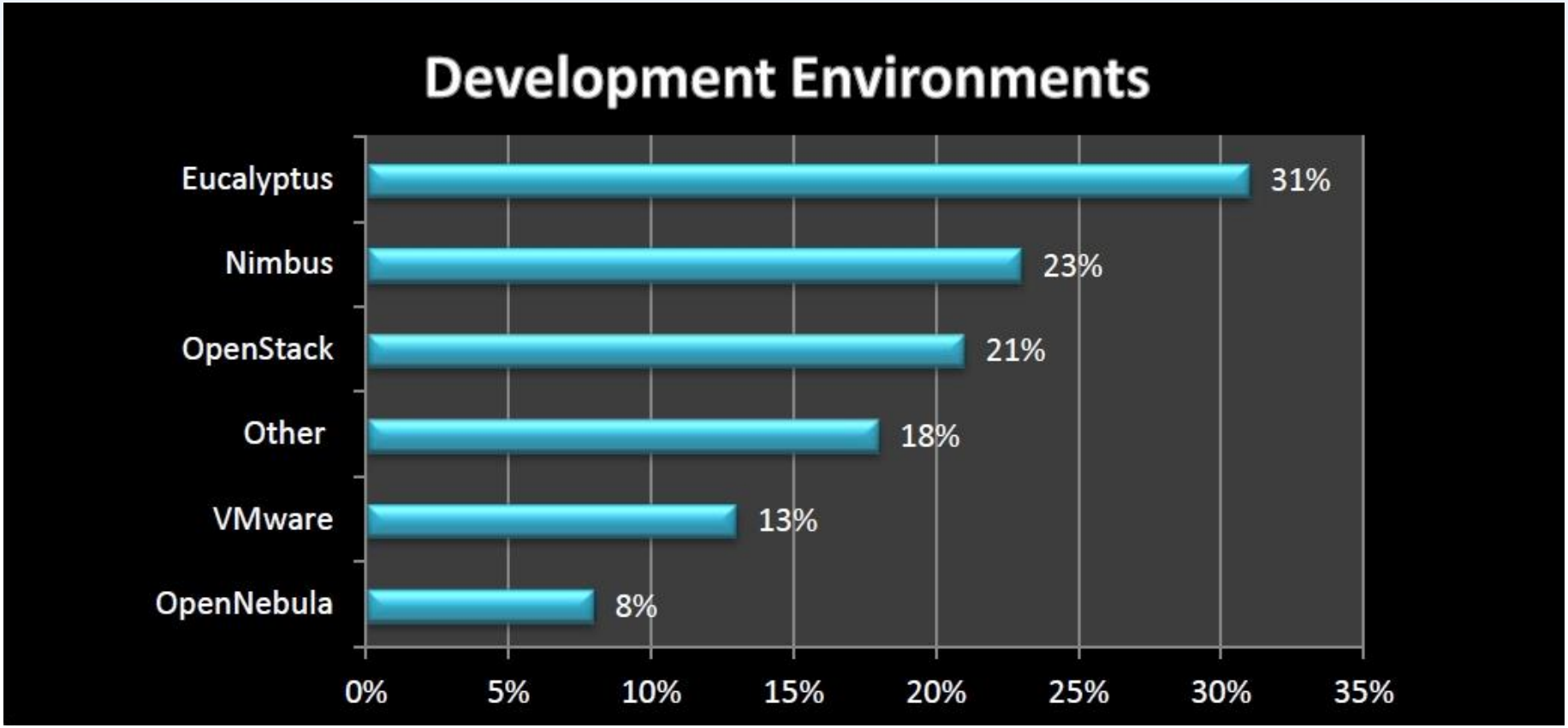
Which service providers did you use? Check all that apply.



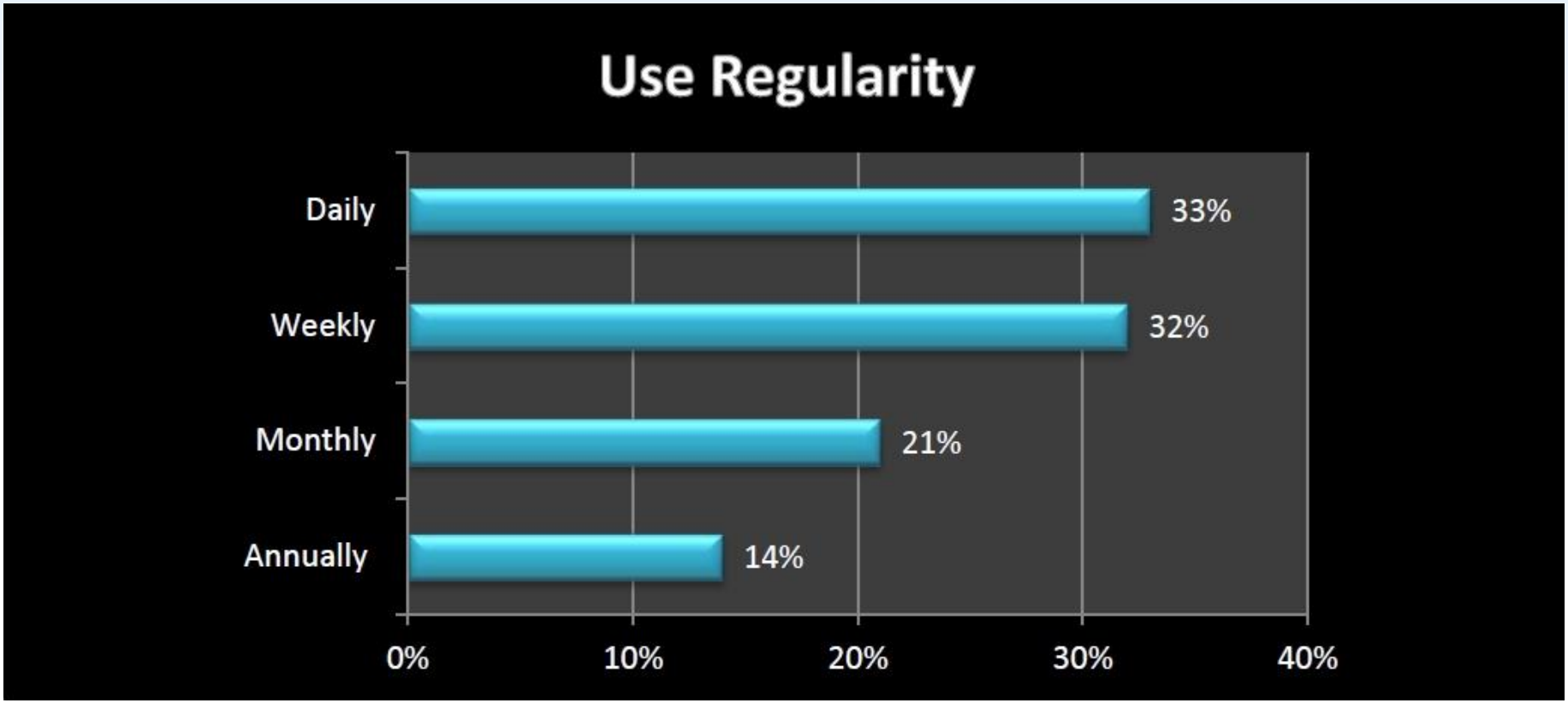
What special features are available from your cloud provider that enabled your research? Check all that apply.



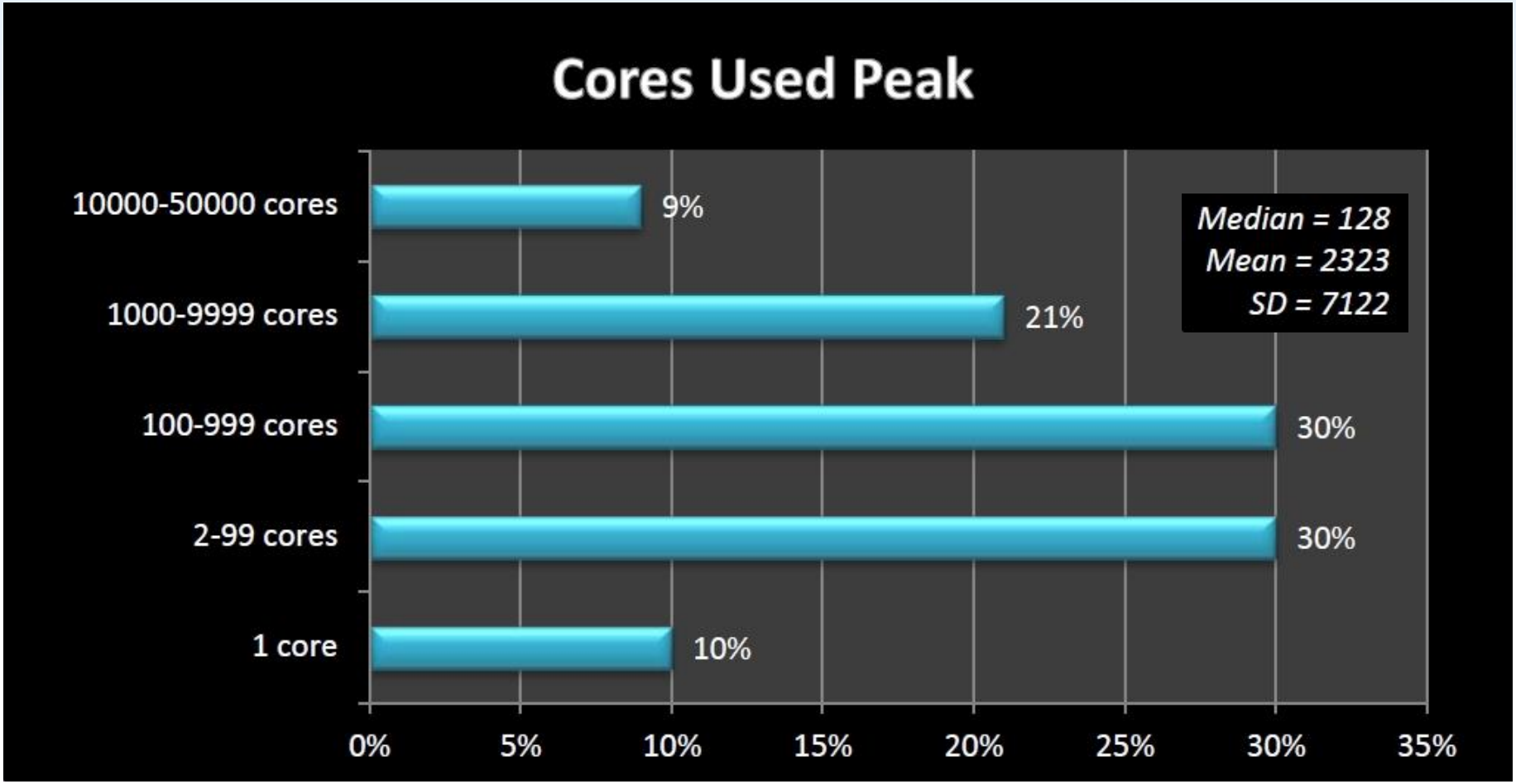
What development environment features available from your cloud provider enabled your research? Check all that apply.



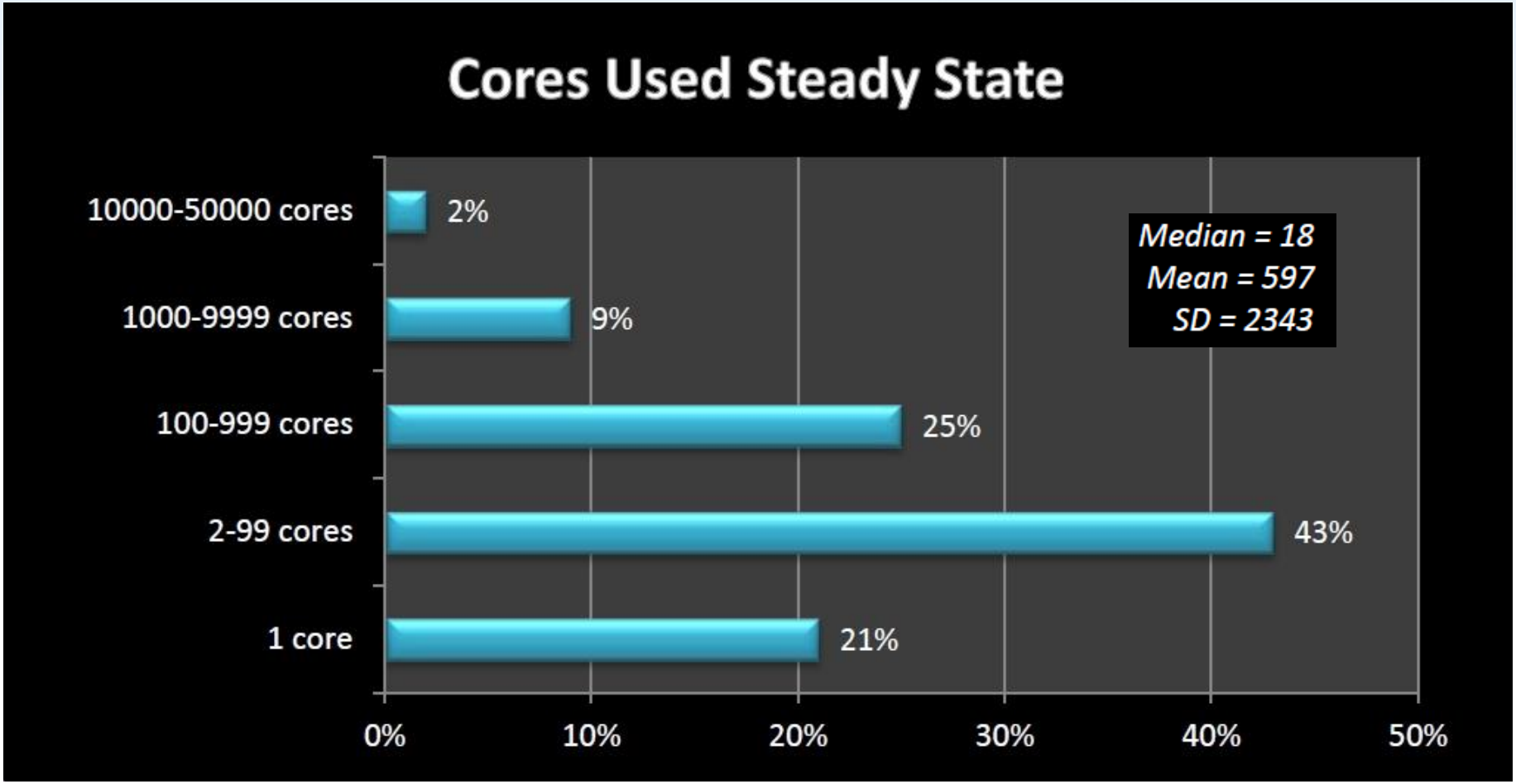
With what regularity do you use cloud resources?



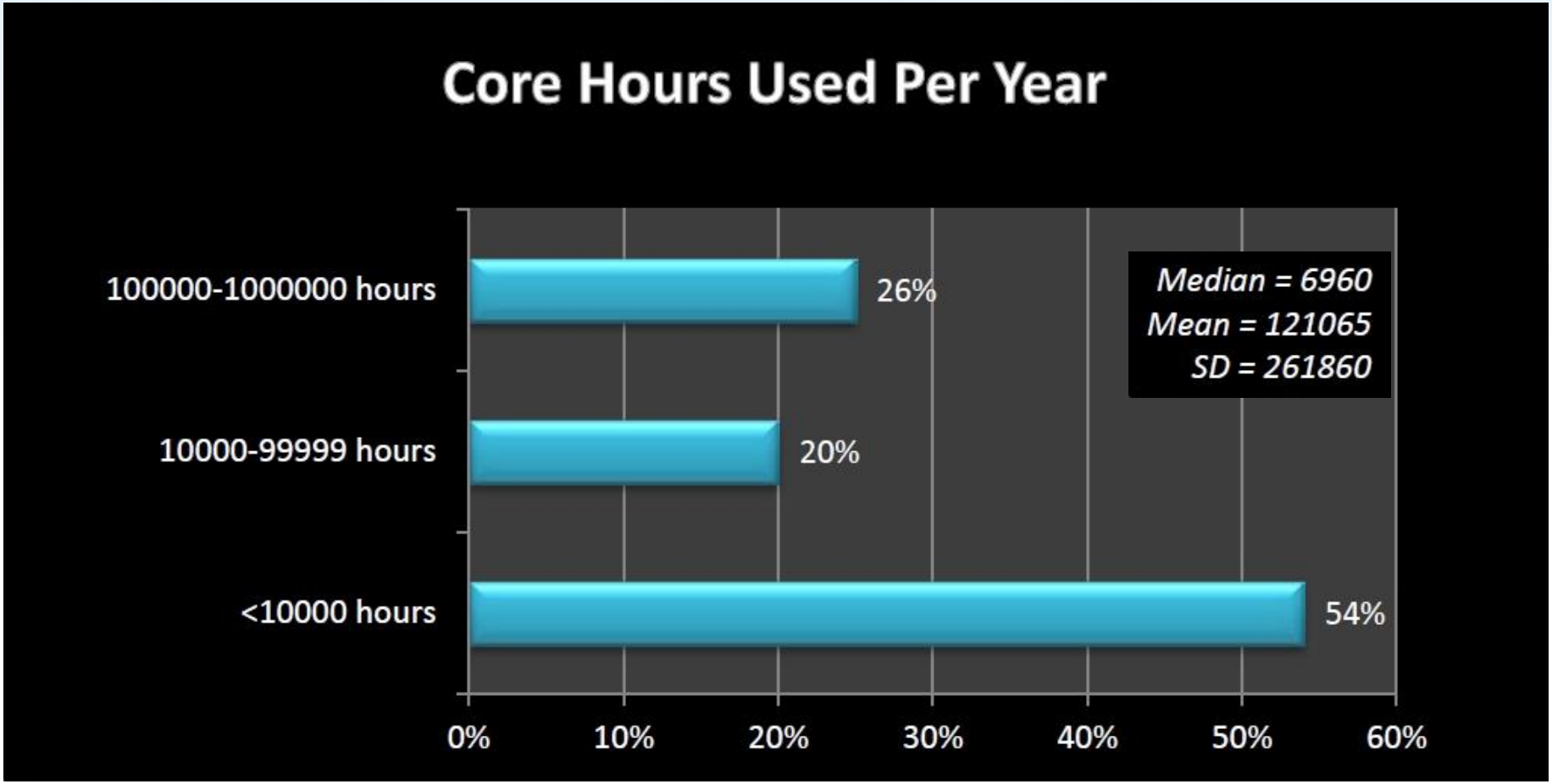
How many cores did you use peak?



How many cores did you use steady state?



How many core hours do you use per year?



Cloud Storage

- 94% access cloud storage for data analysis; 38% archival
- Preferred storage models
  - 27% Object Store, e.g., S3, Swift
  - 23% Elastic Block Storage
  - 21% N/A
  - 15% Other (GlusterFS, RDMS, Window Azure, conventional file systems)
  - 6% Parallel Performance File System
  - 6% HDFS
  - 5% Wide Area File System
- 50% access <100GB during program execution
  - Median accessed during program execution = 80GB
    - Mean = 3.3TB (SD = 12.7TB) due to subset (10%) of very large (10TB-100TB) storage users, e.g., macromolecular modeling
- 55% store <1TB data short-term/long-term
  - Median short-term and long-term storage = 500GB
    - Mean short-term storage = 9.2TB (SD = 56.8TB); mean long-term storage = 7.6TB (SD = 30TB)

User Identified Benefits

1. **Pay as You Go**

"Pay as you go and elasticity are critical."

– Architecture Services CTO

"…cloud enabled the scientific community to access this genome resource quickly without researchers having to procure, deploy, and maintain their own data server."

– Science Gateway Developer

"…you only pay for what you use –when you're not using your 10,000 node Hadoop cluster, you don't pay for it."

– Citizen Science Portal Developer

User Identified Benefits

1. Pay as You Go
2. **Lower Costs**

"…maintenance and administration cost savings are a plus for the cloud."

– Systems Biologist

"…our load CPU demand over a year isn't constant. There are peaks and there are troughs. If we priced our purchase to satisfy our peak needs, we'd find that our system would lay idle for some fraction of the year."

– Particle Physicist

"…no need to purchase an upfront data center for the 5-year mission, as it would be under-utilized most of the time."

– Space Agency Operations Manager

User Identified Benefits

1. Pay as You Go
2. Lower Costs
3. **Compute Elasticity**

"Our 50,000-core compute ran across all 7 Amazon regions using on-demand and spot instances for a computational docking application. The experiment – the equivalent of 12.5 processor years – was conducted in a mere 3 hours. Previously, it would take about 11 days to run a similar analysis on [our] 400-core cluster – stopping all work in the process."

– Software Developer

"We…. reduced our run-time processing for a job analyzing 3.8 million articles from 100 days on our infrastructure down to just 5 days …"

– Data Mining Specialist

User Identified Benefits

1. Pay as You Go
2. Lower Costs
3. Compute Elasticity
4. **Data Elasticity**

"As our platform grows, we anticipate very large datasets to be contributed, so being able to scale quickly is key."

– Supervisor , Energy Science Gateway

"Pay as you go and elasticity are critical. Services such as Amazon Glacier may mean we can leave data in the cloud rather than uploading it every 6 months."

– Astrophysicist

User Identified Benefits

1. Pay as You Go
2. Lower Costs
3. Compute Elasticity
4. Data Elasticity
5. **Software as a Service**

"…the Cloud provides the software we need when we need it, enabling us to develop simulation optimization and feasibility determination algorithms faster and more efficiently. "

– Operations Research Engineer

"…simulations often are too large to execute effectively on desktop workstations (requiring hours to days to weeks to complete), but can be completed in an interactive timeframe (minutes to hours) on Red Cloud with MATLAB. The results then often guide larger scale simulations for which high-end computational resources are absolute necessities."

– *Neuropsychologist*

User Identified Benefits

1. Pay as You Go
2. Lower Costs
3. Compute Elasticity
4. Data Elasticity
5. Software as a Service
6. **Education as a Service**

"We use cloud to address successfully the dual issue of scalability (serving thousands of users at a fairly reasonable quality of service) and sustainability (providing accessibility and availability beyond the classroom)."

– Teaching Tool Developer for freshman biology courses

"I am assembling a collection of open-source tools to support educational development: Calliope for optimization formulations, OCTAVE , and more."

– OR Professor developing online textbook in the cloud

User Identified Benefits

1. Pay as You Go
2. Lower Costs
3. Compute Elasticity
4. Data Elasticity
5. Software as a Service
6. Education as a Service
7. **Broader Use**

"(Our) cloud solution is primarily aimed at domain scientists who do not have advanced IT skills."

– Chemistry Research Associate

"The availability of platform services such as storage and programming abstractions such as .NET or MapReduce reduces the overhead of installing, monitoring and managing such services locally."

– Energy Informatics Director

User Identified Benefits

1. Pay as You Go
2. Lower Costs
3. Compute Elasticity
4. Data Elasticity
5. Software as a Service
6. Education as a Service
7. Broader Use
8. **Scientific Workflows**

"Our projects utilized this new resource to execute scientific workflow applications in a fast, cost efficient way."

– CS Researcher

"For highly performance driven apps that operate on a tightly coupled model, purchasing and managing a rack of ~50 cores is a better model than Cloud resources…. However, much of our research deals with large scale problems rather than high performance problems. In such a scenarios, on-demand access to a large number of virtual machines is more useful than round the clock availability of a captive cluster."

– Director, Energy Informatics

User Identified Benefits

1. Pay as You Go
2. Lower Costs
3. Compute Elasticity
4. Data Elasticity
5. Software as a Service
6. Education as a Service
7. Broader Use
8. Scientific Workflows
9. **Rapid Prototyping**

"The cloud enables us to explore different classes of problems rapidly, opening new doors to research."

– Biological Systems Researcher

"We use the cloud for rapid prototyping. It is also affordable for constant use of small instances for things like MediaWiki and Redmine. Our use is usually data-intensive and access to Red Cloud and GlusterFS avoids the data transfer dilemma."

– IT Director, Life Sciences Core Facility

User Identified Benefits

1. Pay as You Go
2. Lower Costs
3. Compute Elasticity
4. Data Elasticity
5. Software as a Service
6. Education as a Service
7. Broader Use
8. Scientific Workflows
9. Rapid Prototyping
10. **Data Analysis**

"A steep drop in the cost of next-gen sequencing during recent years has made the technology affordable to the majority of researchers, but downstream analysis still poses a resource bottleneck for small labs….We can enable researchers without access to local computing clusters to perform large-scale data analysis, by tapping into a pool of on-demand Cloud BioLinux VMs that can be rented at low cost. Renting servers in the cloud can work as a better model for smaller research labs, where the cost of hardware and maintenance can't be justified for only a few experiments."

– Bioinformatics Engineer

User Identified Challenges

1. **Learning Curve**

"The start-up, programming, and configuration are more challenging than an in-house cluster, however…it isn't difficult to learn."

– Biomechanics Researcher

"The platform may provide the best platform for conducting our research but results are significantly delayed by initial development time."

– Science Gateway Developer

User Identified Challenges

1. Learning Curve
2. **Virtual Machine**

"The virtual machine nature of cloud tends to be detrimental to performance."

– Computational Chemist

"It would be great if the compute instances could be managed in a more flexible and fine-grained manner."

– Computer/Network Security Professor

User Identified Challenges

1. Learning Curve
2. Virtual Machine
3. **Bandwidth**

"…there is no doubt that in the next couple of years we'll see lots of nascent solutions to the fundamental problem of mobility and cloud collaboration: data movement….For a medical computing project, the data set was 35TB, but the data expansion of these data sets could be as high as 100GB per day, fueled by high volume instruments such as MRI or NGS machines. In the US-China collaboration, the problem was latency and packet loss, whereas in the medical project, it was how to deal with multi-site high-volume data expansions."

– Global Engineering Consultant

User Identified Challenges

1. Learning Curve
2. Virtual Machine
3. Bandwidth
4. **Memory Limits**

"RAM limitations – I need more than the maximum provided by Amazon (and most cloud providers). 300GB+ needed."

– Molecular Genetics Researcher

"The configurations are fixed so sometimes we waste memory or CPU."

– Astrophysicist

User Identified Challenges

1. Learning Curve
2. Virtual Machine
3. Bandwidth
4. Memory Limits
5. **Databases**

"The cloud is less stable than a local server or HPC machines and may shut down unexpectedly because of upgrades or because some unmanaged exception in other processes, plus non-relational DBs require architecting and coding effort to ensure transactional operations in order to preserve consistency – your code may be shut down at any minute."

– Biological Systems Researcher

User Identified Challenges

1. Learning Curve
2. Virtual Machine
3. Bandwidth
4. Memory Limits
5. Databases
6. **Interoperability**

"The time-critical nature and dynamic computational workloads of Value at Risk (VaR) applications make it essential for computing infrastructures to handle bursts in computing and storage resources needs….integrating clouds with computing platforms and data centers, as well as developing and managing applications to utilize the platform remains a challenge."

– Software Developer

User Identified Challenges

1. Learning Curve
2. Virtual Machine
3. Bandwidth
4. Memory Limits
5. Database
6. Interoperability
7. **Security**

"The issues (our ethical hacker) found were almost entirely challenges we would face and issues we would have to protect against whether this was locally hosted, using our on premise physical infrastructure, or remotely hosted at a public cloud provider….It is reasonable to assume further efforts may be needed if a higher level of isolation is demanded for specific confidential data. However, our results affirmed our belief that institutions such as our own can responsibly utilize cloud and public cloud providers."

– Senior Fellow, Inter University Consortium for Political and Social Research

User Identified Challenges

1. Learning Curve
2. Virtual Machine
3. Bandwidth
4. Memory Limits
5. Databases
6. Interoperability
7. Security
8. **Data Movement**

"Most of our collaborators have the following view of cloud resources: clouds are excellent at providing burst capability and custom software environments for computation and data analytics….On the storage side, they are very concerned about the high cost of long-term storage, and the risk of data loss or extreme cost retrieval. They are much more comfortable keeping their data at the local campus, where they can control and access it on demand."

– Software Researcher and Designer

User Identified Challenges

1. Learning Curve
2. Virtual Machine
3. Bandwidth
4. Memory Limits
5. Databases
6. Interoperability
7. Security
8. Data Movement
9. **Storage**

"I wish EBS volumes would work more like Lustre file systems, i.e., high performance, high availability, and the ability for multiple VMs to read/write to one EBS volume."

– Bioinformatics Researcher

"Due to our application requirements, we'd like the cloud to provide secure data store and allow users to customize and copy their VM instances."

– Academic/Industry Research Collaboration

User Identified Challenges

1. Learning Curve
2. Virtual Machine
3. Bandwidth
4. Memory Limits
5. Databases
6. Interoperability
7. Security
8. Data Movement
9. Storage
10. **Cost/Funding**

"Opacity of cost is a problem. We were occasionally surprised by how much we were spending on certain resources."

– EE Postdoc

"We find it difficult to write cloud compute resources into our grants."

– Citizen Science Director

"Justifying paying for cloud resources on NSF grants…"

– CS Prof

"Overall cost and charging to a grant that does not have such a cost model built in are challenges."

– Regional Data Center Researcher

Closing Thoughts

- Clouds have a ***complementary role*** to play in research infrastructure portfolios
- Clouds ***deliver value*** to researchers and educators.
  - scaling high throughput, non-tightly coupled applications on-demand
  - data management and analysis
  - Software as a Service (SaaS)
  - access to discipline-specific gateways with optimized workflows
  - on-demand labs, digital textbooks and other interactive learning experiences (EaaS)
- Clouds ***mask computing complexities*** for non-HPC/non-IT savvy users
- Clouds ***democratize access*** for institutions who are not resource rich
  - Benefiting greater number/greater diversity of scientists and HASS researchers
- ***Production resource clouds needed*** as part of national/regional research infrastructure (SaaS, discipline-specific clouds, etc.) and/or ***seamless access to public clouds*** as part of research grant funding
- ***Cloud user training and consulting needed*** to reduce learning curve
- Continued ***CS research needed*** to address cloud challenges/limitations


Our reach will forever
exceed our grasp, but,
in stretching our horizon,
we forever improve our world.
XSEDE
Extreme Science and Engineering
Discovery Environment